Identifying genes of gene regulatory networks using formal concept analysis

Jutta Gebert*[a], Susanne Motameny*[a],
Ulrich Faigle[a], Christian V. Forst[b], Rainer Schrader[a]

[a]Center for Applied Computer Science, University of Cologne, Weyertal 80, 50931 Cologne, Germany

[b]Los Alamos National Laboratory, PO Box 1663, Mailstop M888, Los Alamos, NM 87545, USA

May 2, 2007

Abstract

Motivation:
In order to understand the behavior of a gene regulatory network, it is essential to know the genes that belong to it. Identifying the correct members (e.g. in order to build a model) is a difficult task even for small subnetworks. Usually only few members of a network are known and one needs to guess the missing members based on experience or informed speculation. It is beneficial if one can additionally rely on experimental data to support this guess. In this work we present a new method based on formal concept analysis to detect unknown members of a gene regulatory network from gene expression time series data.

Results:
We show that formal concept analysis is able to find a list of candidate genes for inclusion into a partially known basic network. This list can then be reduced by a statistical analysis so that the resulting genes interact strongly with the basic network and therefore should be included when modeling the network. The method has been applied to the DNA repair system of *Mycobacterium tuberculosis*. In this application our method produces comparable results to an already existing method of component selection while it is applicable to a broader range of problems.

Contact: gebert@zpr.uni-koeln.de, motameny@zpr.uni-koeln.de

*Both authors contributed equally to this work

1 Introduction

Gene expression data is one of the most often analyzed kind of data in the field of bioinformatics as it serves as a basis to answer several different questions. One of the goals is the understanding of processes in the cell on a molecular level. Time series measurements of gene expression levels can reveal the different regulatory interactions between genes. A gene regulatory network is defined as a graph with genes as nodes and their interactions as edges. Often functions are assigned to the edges to determine the dynamic behavior of the concentrations of the corresponding mRNAs (see for example [Gebert *et al.*, 2006, Radde *et al.*, 2006]). In most studies only subgraphs representing a small subnetwork and the corresponding interactions are examined. In order to build a reasonable model for such a subnetwork, it is essential that all components of the network are identified.

Usually, some genes are known to be responsible for most of the behavior of the subnetwork. Their mutual influences might be known to some extent as well. We will call such genes *seed genes* hereafter. Now the question is: How to determine possible other members of the network when gene expression time series data are available?

In this paper we will present a method which finds genes that might have an important influence on the gene regulatory network defined by the seed genes and that should be added to this network. A mathematical method, called formal concept analysis, is used to detect a list of candidate genes that can be analyzed afterwards in the same way as proposed in [Radde *et al.*, 2006]. In that paper a graph theoretical approach was used to detect such a list of genes, which is further described in [Cabusora *et al.*, 2005]. The graph theoretical approach needs interaction data as well as expression data to find k-shortest paths in the graph built from the interaction data. In our new approach the availability of interaction data is not a prerequisite. Nevertheless, if such data are available they are included in the preprocessing step of our approach based on formal concepts and statistical analysis.

Another method concerning the choice of components in gene regulatory networks has been proposed by [Hashimoto *et al.*, 2004]. They use gene expression data to grow a gene regulatory network out of a seed consisting of one or more genes and apply their method to a glioma gene expression data set. Iteratively new genes are adjoined to the seed so that the autonomy of the network is enhanced. Autonomy is defined with the help of two properties of subnetworks: The genes should interact significantly and they should not be strongly influenced by genes outside the network. Later, another approach for growing a subnetwork around a gene of interest was proposed by [Bansal *et al.*, 2006]. They perturbed the gene of interest and measured the following gene expression profiles at several time points to infer gene regulatory networks.

Instead of growing a regulatory network from few genes the opposite approach is also possible. Standard clustering algorithms find genes that are co-regulated. However, there are two problems. Genes are usually only assigned to one cluster although they might play a role in more than one regulatory module, and

they might be active in only few experiments so that the noise level is very high (see [Kloster *et al.*, 2005]). Signature algorithms as the one proposed by [Ihmels *et al.*, 2002] can overcome these problems. This algorithm determines a set of co-regulated genes together with conditions under which they are co-regulated. An extension of the algorithm which uses the knowledge of already identified modules in each iteration step can be found in [Kloster *et al.*, 2005].
Our method is based on formal concept analysis and is applied to the DNA repair system of *Mycobacterium tuberculosis*. The method proposed in [Radde *et al.*, 2006] uses the same data set so that the methods can be compared with each other.
The bacterium *M. tuberculosis* has been discovered by Robert Koch in 1882. Its genome was sequenced in 2000, thus enabling the understanding of processes in its cell on molecular level. Detailed knowledge about regulatory mechanisms can build the basis for drug intervention. One of these regulatory mechanisms is the DNA repair system. The system consists of the proteins LexA and RecA as well as up to 40 genes that are regulated by these two proteins together. If a damage in the DNA occurs which results in single-stranded DNA, then the DNA repair system is activated. The protein RecA binds to this single-stranded DNA and produces a change of the structure of the protein LexA. In this form LexA cannot bind to the regulatory regions called SOS boxes on the DNA any longer. The associated SOS genes will be no longer repressed by LexA and transcription starts. Both genes *recA* and *lexA* have an SOS box as well. Additionally to the SOS regulation there exists another mechanism in *M. tuberculosis* to activate some of the SOS genes in the case of DNA damage (see [Dullaghan *et al.*, 2002]). In [Radde *et al.*, 2006] the gene *Rv2719c* is predicted to play an important role in the DNA repair system. We will compare our results with this prediction.
In section 'Methods' we will present the analysis of gene expression data and interaction data with formal concept analysis followed by the statistical analysis proposed by [Radde *et al.*, 2006]. In section 'Application' we apply this method to data of *M. tuberculosis* and in section 'Discussion' we will compare the results with the results given in [Radde *et al.*, 2006] and discuss the prospectives of our approach. The section 'Conclusion' will give a short summary of this paper.

2 Methods

We assume that $G = \{g_1, ..., g_n\}$ is the set of all genes of an organism and a subset $S \subseteq G$ is our set of seed genes. We want to find another subset $\tilde{S} \subseteq G \setminus S$ of genes that interact strongly with the network defined by S. Moreover, we want to determine the structure of the new network consisting of genes S and $\tilde{S}$. This means, we want to find pairs of genes that influence each other.
Let $R \subseteq G \times G$ be a relation that contains interaction data and M an $n \times l$ matrix consisting of gene expression time series data of length l. If $(g_i, g_j) \in R$ we know that g_i and g_j interact with each other, for example g_i inhibits g_j or g_j activates g_i. The kind and direction of interaction is not given here. The matrix M consists of the entries m_{ij} which is the expression of gene i at time point j in the experiment. The rows of M are vectors of the form $m_i = (m_{i1}, \ldots, m_{il})$

and describe the expression of gene i at time points $1, \ldots, l$. We will use the term *expression profile* to refer to such a vector m_i throughout the paper.
We will proceed in three steps:
1. In a preprocessing step we use the relation R to get a first list of interesting genes. If no interaction data are available, this step is skipped and we start directly with the second step using the whole gene set G as the first list.
2. Now we construct a concept lattice from the gene expression data to reduce the number of genes on the first list.
3. In the last step we calculate probabilities for the correlation coefficients between the genes that result from the second step and genes of S in order to obtain significant interactions.

2.1 Preprocessing step

In a preprocessing step the relation $R \subseteq G \times G$ is used to reduce the set G. If such interaction data are not given, this step can be left out and we can start directly with the second step. We define gene lists S_k by including genes into the list that are related to the seed genes due to R, and by including also the related genes to these included genes and so on:

$$\begin{aligned} S_1 &:= \{g \in G : \exists s \in S \text{ with } (s,g) \in R\} \\ S_k &:= S_{k-1} \cup \{g \in G : \exists s \in S_{k-1} \text{ with } (s,g) \in R\} \text{ for } k \geqslant 2. \end{aligned}$$

Let us assume we represent the interaction data by a graph. The nodes are the genes and an edge (g_i, g_j) is drawn whenever $(g_i, g_j) \in R$. Then the sets S_k can be interpreted with the help of such an interaction graph as follows. The set S_k consists of all genes that are contained in a path with maximum length k so that this path starts at one of the seed genes. The sets will be calculated for all $k \leqslant K$ where $K \in \mathbb{N}$ is a given threshold. The smaller k the less genes we will get as a result in the set S_k. Thus the parameter K can be used for scaling the network. The smaller K the fewer genes will be included into the network (see also figure 1).

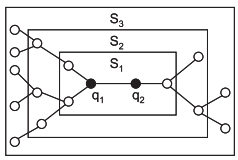

Figure 1: In this example the genes q_1 and q_2 are the seed genes and are marked by black circles in the interaction graph. The rectangles in the figure include the corresponding gene lists S_1, S_2 and S_3.

2.2 Template Matching

For the determination of the candidate genes that are potentially related to the basic network, we first characterize the gene expression profiles by means of template matching. In order to do this, m templates $\text{Temp}_1, \ldots, \text{Temp}_m$ are defined, where each template is a real-valued vector of the same length as the expression profiles. As a measure of similarity between a gene expression profile and a template, a noise-robust Kendall correlation coefficient is used.
Given a noise level $\rho \geqslant 0$ and two time series $X = (X_1, ..., X_l)$ and $Y = (Y_1, ..., Y_l) \in \mathbb{R}^l$ with non-constant entries, we compare every pair (X_i, Y_i) to each other pair (X_j, Y_j). If

$$(X_i > X_j + \rho) \wedge (Y_i > Y_j + \rho) \text{ or } (X_i < X_j - \rho) \wedge (Y_i < Y_j - \rho)$$

holds, then (X_i, Y_i) and (X_j, Y_j) form a *proversion*. An *inversion* is given whenever

$$(X_i > X_j + \rho) \wedge (Y_i < Y_j - \rho) \text{ or } (X_i < X_j - \rho) \wedge (Y_i > Y_j + \rho).$$

A *binding* in X is given if

$$X_i - \rho \leqslant X_j \leqslant X_i + \rho$$

and a binding in Y is defined analogously. Based on these notions the noise-robust Kendall correlation coefficient is defined by

$$\tau_K(X, Y) = \frac{P - I}{\sqrt{(\frac{l(l-1)}{2} - B_X)(\frac{l(l-1)}{2} - B_Y)}}, \tag{1}$$

where P is the number of proversions, I is the number of inversions, B_X is the number of bindings in X and B_Y is the number of bindings in Y. If no proversion or inversion is found, the coefficient is set equal to zero. In the special case that $\rho = 0$, the above definition corresponds to the classical Kendall correlation coefficient. The introduction of the noise level ρ makes it applicable to gene expression time series data, which are relatively noisy due to several sources of error.
The Kendall correlation coefficient not only detects a linear relationship between the time series as the commonly used Pearson correlation coefficient but also accounts for increasing or decreasing monotone functions. Another advantage relates to outliers which do not have such a strong effect on the proposed coefficient as they have on the Pearson correlation coefficient.

We say that a gene *matches* a template if the absolute value of the noise-robust Kendall correlation coefficient exceeds a certain threshold t_{match}. The absolute value is chosen because we want to include genes that show parallel regulation (i.e. are up-regulated along with a seed gene) as well as genes which show opposite regulation (i.e. are down-regulated while a seed gene is up-regulated) in our candidate list. The threshold t_{match} is chosen in such a way that each gene can match several templates and for the next step, the construction of the concept lattice, we assume that each gene expression profile is characterized by the templates it matches.

2.3 Constructing the concept lattice

Concept lattices and *formal concept analysis* were introduced in [Wille, 1982] (see also [Ganter and Wille, 1996]). A concept lattice exposes the logical structure of object-attribute data. The objects we will consider are the genes and the attributes are the templates from the template matching procedure. As stated in the previous paragraph, a gene g_i matches the template Temp_j if the noise-robust Kendall correlation of its expression profile $m_i = (m_{i1}, ..., m_{il})$ with the pre-defined template Temp_j exceeds the threshold t_{match}. It is important to set the threshold in such a way that genes can match several patterns. Otherwise, the concept lattice will yield no interesting information. In general, the object-attribute data can be represented in a table with the objects as row indices and the attributes as column indices and a cross at the entry (i, j) if object i possesses the attribute j. Such a table that contains crosses and empty entries is called a *formal context*. In our scenario, we get a formal context with n rows (one for each gene) and m columns (one for each template) and a cross in entry (i, j) if gene g_i matches the template Temp_j.
Next we introduce the notion of a formal concept. By a formal concept we understand a maximal set of genes A in which every gene matches the same templates (contained in the set B), together with these templates. We mathematically define a formal concept as follows:

Let A be a subset of genes and B be a subset of templates. We define

$$A' := \{\text{templates Temp} \mid g \text{ matches Temp } \forall g \in A\} \tag{2}$$

$$B' := \{\text{genes } g \mid g \text{ matches Temp } \forall \text{ Temp} \in B\}. \tag{3}$$

A pair (A, B), consisting of a set of genes A and a set of templates B, is called a *formal concept* if $A' = B$ and $B' = A$. A is called *extent* and B is called *intent* of the concept (A, B). The formal concepts are ordered according to set-inclusion on the extents:

$$(A_1, B_1) \leqslant (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$$

The interpretation is simple: If we increase the number of genes belonging to a concept, the number of jointly matched templates will decrease – we obtain a *superconcept* of the initial one. A *subconcept* is obtained by decreasing the number of genes, so that the remaining genes simultaneously match more templates. The hierarchy of concepts induced by the $\leqslant$-order is called a *concept lattice* and exhibits the logical structure of the context table. In the concept lattice, genes belonging to one concept share certain characteristics of their expression profiles. Genes belonging to concepts near the bottom of the lattice share many characteristics, whereas genes near the top share few characteristics.
Candidate genes for inclusion into the network are, roughly speaking, those genes that behave similar to the seed genes. Genes behave similar if they match similar sets of templates. A candidate gene for the network should match the

templates of a seed gene and should not match templates that are not matched by seed genes. In the language of formal concept analysis, candidate genes belong to concepts whose intents are subsets as well as supersets of intents corresponding to seed genes. Let (A_1, B_1) and (A_2, B_2) be concepts that contain seed genes in their extents and where $B_2 \subset B_1$ (i.e. (A_1, B_1) is a subconcept of (A_2, B_2)). Then all candidate genes are contained in the extents of concepts (A, B) which satisfy $(A_1, B_1) \leqslant (A, B) \leqslant (A_2, B_2)$. This candidate list is then further narrowed down by statistical analysis.

2.4 Statistical analysis

The candidate list includes the seed genes S as well as some other genes S^*. We calculate the noise-robust Kendall correlation coefficient between the expression profiles of genes in S and those of genes in S^*.

All correlation coefficients between the seed genes and all other genes in G are now calculated. They form an underlying distribution for the following statistical approach. Let $K = (S \cup S^*, E)$ be an undirected complete bipartite graph, with vertex set $S \cup S^*$ and undirected edges E existing between each pair of vertices (v_1, v_2) with $v_1 \in S$, $v_2 \in S^*$. Given a significance level $\alpha > 0$ we compute $Q_{\min}$ as the maximum of the $(\alpha/2)\%$ smallest correlation coefficients between the seed genes and genes in G. Similarly, $Q_{\max}$ is obtained as the minimum of the $(\alpha/2)\%$ largest of these values (see also [Radde *et al.*, 2006]). We define an edge to be significant with respect to a given α if the correlation coefficient between the two expression profiles of the corresponding genes is smaller than $Q_{\min}$ or greater than $Q_{\max}$. Next we delete all insignificant edges from K and also those vertices which are no longer connected to the seed genes S. With this procedure we obtain a graph that represents strong interactions between genes of S^* and S. As the seed genes are assumed to be well-known, interactions between the seed genes are added in the graph on the basis of such biological knowledge. Given that the underlying seed genes S are connected themselves, the obtained graph is connected, too. Again biological knowledge has to be included to find the directions of regulations. In the section 'Application' this is achieved for example for some genes by including the knowledge of whether these genes have an SOS box or not.

Remark: Short time series as the one used in the application or time series which have large time gaps in between the measurements will be analyzed in our method with the noise-robust Kendall correlation coefficient. A large coefficient could result from a co-regulation of these genes or from one gene regulating the other. Long time series with short time gaps between the measurements can also exhibit the direction of regulation. Such a time dependent response can be detected with the shifted noise-robust Kendall correlation coefficient. It is defined for each pair (X, Y) by

$$\tau_{K,t^*}(X, Y) \quad = \quad \tau_K(X^*, Y^*) \tag{4}$$

with $X^* = (X_1, ..., X_{l-1})$ and $Y^* = (Y_2, ..., Y_l)$. Here, $\tau_{K,t^*}(X, Y)$ will have a high absolute value, if the time series Y shows the same pattern as time series

X but shifted one time point to the right. The shift can also include several time steps to allow for delayed regulations.
Using the shifted noise-robust Kendall correlation coefficient we assume that we have a directed complete bipartite graph $K = (S \cup S^*, E)$, which is defined as having vertices consisting of the set $S \cup S^*$ and directed edges existing from v_1 to v_2 and from v_2 to v_1 for each $v_1 \in S$, $v_2 \in S^*$. Defining significant edges with respect to a given $\alpha > 0$, deleting edges and vertices in K in the same way as before and defining the interactions between the seed genes due to biological knowledge, we get as a result a directed regulatory graph.
The genes of the graph, that are not the seed genes, but remain in the graph after the deletion of the edges and vertices, form the set $\tilde{S}$ which we defined as the set of genes that strongly interact with our seed genes.

3 Application

3.1 SOS repair system of *Mycobacterium tuberculosis*

The gene expression data from *M. tuberculosis* comprise measurements from 8 time points for mitomycin experiments, and for some of these time points, replicates are available. The experiments have been conducted by [Boshoff *et al.*, 2004] and are available at the NCBI 'Gene Expression Omnibus'. As a first step, preliminary analyses were performed using BRB ArrayTools developed by Dr. Richard Simon and Amy Peng Lam. These included the normalization of the data and the identification of bad measurements resulting in missing values and removal of genes. Then, in order to generate an 8-time-point profile for each gene, the mean of all replicates for each time point was taken. If for one gene all of the measurements at one time point were missing, the resulting mean value was also marked as a missing value in the gene's profile. Finally, missing values in the profiles were estimated by a nearest neighbor approach (see e.g. [Troyanskaya *et al.*, 2001]).
As the set of seed genes we define S=$\{$*recA, lexA, ruvC, linB*$\}$. The genes *recA* and *lexA* are two of the main components in the DNA repair system. They together regulate all SOS genes. As we try to get insight into the unknown alternative mechanism we include one gene that is solely regulated by the *recA-lexA*-mechanism, namely *linB*, and one gene, which is additionally regulated by an alternative mechanism, namely *ruvC*. This set of seed genes coincides with the set of seed genes chosen in [Radde *et al.*, 2006] so that the approaches are comparable.
The preprocessing step is carried out with interaction data about *M. tuberculosis*. The parameter $K = 4$ is chosen, which results in a list of 12 genes in S_1, 196 genes in S_2, 808 genes in S_3 and 925 genes in S_4. These lists were reduced in several filtering steps which include the removal of genes that have many missing values and the removal of genes which do not show at least a 2-fold change in their gene expression levels for at least one time point. After the filtering, 8

genes remain in S_1, 68 in S_2, 258 in S_3, and 339 in S_4.
Six templates are defined which capture the qualitative behavior of the seed genes observed in the experiment. They are given by

$$\begin{aligned}
\text{Temp}_1 &= (0,0,0,1,2,3,2,2)\\
\text{Temp}_2 &= (0,0,0,0,1,2,1,1)\\
\text{Temp}_3 &= (0,0,0,1,2,3,2,1)\\
\text{Temp}_4 &= (0,0,0,0,1,2,1,0)\\
\text{Temp}_5 &= (0,0,0,1,2,3,3,3)\\
\text{Temp}_6 &= (0,0,0,1,1,2,1,1).
\end{aligned}$$

The threshold t_{match} for the template matching procedure is set to 0.6 so that genes match several templates. The resulting concept lattices contain 7 concepts for S_1, 24 concepts for S_2, 26 concepts for S_3 and 28 concepts for S_4. A list of 3, 13, 55 and 65 candidate genes, respectively, is defined by the concepts between the seed genes. Using ToscanaJ, a software described in [Becker *et al.*, 2002], the concept lattice for S_2 is generated and visualized as a line diagram (see figure 2). The line diagram is a labeled Hasse-diagram of the hierarchy of concepts. Each concept is represented by a circle and a superconcept always appears above all of its subconcepts. Two concepts are connected by a line if the one is a direct subconcept of the other. There are two types of labels that are attached to the circles. The label above the circle contains attributes (we will use the term *attribute-label*) whereas the label below the circle lists objects (*object-label*). As each concept contains all attributes of its superconcepts as well as all objects of its subconcepts, it is sufficient to label a concept with an attribute if it is the greatest concept containing this attribute. Dually, an object is listed only in the label of the smallest concept to which it belongs. With this labeling scheme, extent and intent of any concept can be read from the line diagram: In order to determine the extent of the concept, one follows all descending paths from the concept and collects all objects that are listed in object-labels along these paths. The intent is found by following all ascending paths in the diagram and collecting all attributes in the same way. The concept with the object-label *lprJ* e.g. has the extent {*lprJ, linB, uvrA*} and intent {Temp1, Temp2, Temp4, Temp5}.
Table 1 shows the probabilities of the noise-robust Kendall correlation coefficient between the seed genes and the genes of the candidate list generated from S_2. Deleting the insignificant edges with respect to $\alpha = 0.05$ in the undirected complete bipartite graph K, we obtain the set $\tilde{S}$ consisting of 5 genes.

Therefore, the seed genes and these five genes *Rv2719c, dnaB, dnaE2, fadD21* and *fadD23* are kept in the gene regulatory graph. The result for S_2 is shown in figure 3. The interaction between the seed genes is known and therefore drawn in the figure using directed edges. Edges between seed genes and all other genes are according to table 1. A significant value will result in an undirected edge.

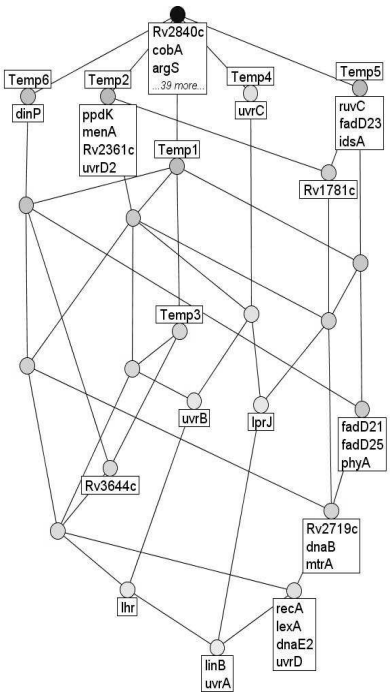

Figure 2: Specifying S_2 according to the preprocessing step, the lattice shown in this figure is built according to gene expression data for 8 time points in mitomycin experiments.

Choosing S_3 instead of S_2 the genes $\{$*Rv0059, Rv0881, Rv2165c, accD2, ald, alkA, fadE21, fbpC2, fmt, ltp1, sdhA, thiL*$\}$ join the set $\tilde{S}$, which then consists of 17 genes.
For S_4, $\tilde{S}$ is further extended by the gene $\{$*rpe*$\}$. Setting K higher and higher or alternatively taking the whole set G instead of a set S_k, the set $\tilde{S}$ will finally include all genes whose expression profiles are close to the seed genes in the sense of formal concept analysis and have very high correlation coefficients to at least one of the seed genes.

4 Discussion

Five genes are added to the subnetwork consisting of the seed genes *recA, lexA, ruvC* and *linB*. The function of most of these genes is already known or conjectured.

genes	*lexA*	*recA*	*ruvC*	*linB*
Rv1781c	0.498	0.476	0.643	0.428
Rv2719c	0.118	0.051	0.204	**0.026**
dnaB	**0.043**	**0.031**	**0.043**	**0.022**
dnaE2	0.051	**0.023**	**0.011**	**0.023**
fadD21	0.164	**0.033**	**0.039**	0.065
fadD23	**0.033**	**0.046**	**0.046**	0.260
fadD25	0.139	0.067	0.170	0.129
idsA	0.152	0.115	0.080	0.303
lprJ	0.297	0.225	0.363	0.081
mtrA	0.462	0.266	0.320	0.141
phyA	0.164	0.097	0.087	0.156
uvrA	0.191	0.129	0.261	0.068
uvrD	0.350	0.278	0.436	0.188

Table 1: Probabilities to get the beforehand calculated noise-robust Kendall correlation coefficient or an even higher deviation from the mean. Bold faced values are significant values.

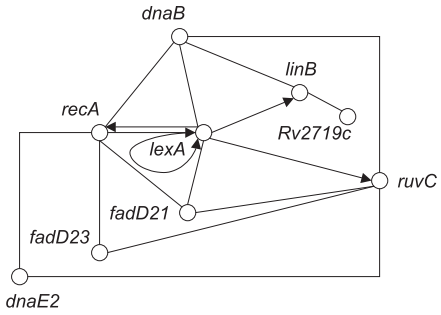

Figure 3: Gene regulatory network based on S_2.

- *Rv2719c*: This gene has been shown to be a DNA-damage inducible gene and is thought to be regulated by LexA (see [Dullaghan *et al.*, 2002]).
- *dnaE2*: It probably encodes a DNA polymerase, which is a complex enzyme responsible for most of the replicative synthesis in bacteria.
- *dnaB*: It probably encodes a replicative DNA helicase.
- *fadD21*: It probably encodes a fatty-acid-CoA ligase.
- *fadD23*: It also probably encodes a fatty-acid-CoA ligase.

The information about the gene functions are listed in the genomic database TubercuList[1] of the Pasteur Institute in Paris. One of these genes, *Rv2719c*,

[1] http://genolist.pasteur.fr/TubercuList/

has been discovered to play an important role in the DNA repair system before (see [Dullaghan *et al.*, 2002, Radde *et al.*, 2006]). *Rv2719c* is the gene that was included in the model of the DNA repair system by [Radde *et al.*, 2006], which resulted in an improvement of the simulations compared to the first model without that gene. The genes *dnaE2* and *dnaB* are also involved in DNA repair or replication and therefore seem to be a biologically reasonable result. The last two genes *fadD21* and *fadD23* have both relatively different functions in contrast to the other three genes. Also, their expression profiles show a down-regulation, while the other three candidate genes are up-regulated along with the seed genes.
The proposed method based on formal concept analysis has two advantages compared to the method suggested in [Radde *et al.*, 2006]. Interaction data may be used but they are not a prerequisite. Only in few cases interaction data are available for the whole genome so that applications are limited if interaction data are necessary for the method. Moreover, the concept lattice gives a good overview of the correlation dependent structure between gene expression profiles if the templates are constructed thoroughly. The combination of template matching and formal concept analysis is better than working with the correlations between seed genes and other genes directly for a decisive reason: It captures the qualitative behavior of the expression profiles while ignoring much of the variation introduced by noise and outliers. We thus can take a step back and observe the big picture instead of getting lost in disorienting details.
In this paper, we use formal concept analysis to structure a set of genes according to the characteristics of gene expression profiles. As concept lattices can be built on any kind of object-attribute data, numerous other applications of this method to biological data sets are imaginable. To extend the use of formal concept analysis on biological problems, further research is underway. The key issue here will be to make formal concept analysis applicable to large data sets as they are produced in the molecular biological sciences and to provide a flexible and informative visualization of the produced concept lattices.

5 Conclusion

We propose a new method to find genes that should be included into regulatory network models that uses formal concept analysis. The method uses gene expression time series and a set of seed genes which are known to be important components of the network. A correlation based characterization of the expression profiles is used to produce a concept lattice from which the candidate genes for inclusion into the network can be extracted. Further refinement of the list is achieved by a subsequent statistical analysis which only keeps those candidate genes that show a statistically significant correlation to the expression profiles of the seed genes. We apply the method to the DNA repair system of *M. tuberculosis* with 4 seed genes and identify a list of 5 genes that should be included into the network model. The majority of these genes are known or suspected to play a role in the DNA repair system.

Acknowledgment

Jutta Gebert was supported by BMBF and Susanne Motameny was partly supported by BMBF, the Rectorate of the University of Cologne and the Koeln Fortune Program of the Faculty of Medicine.

References

[Bansal *et al.*, 2006] Bansal, M., Gatta, G.D., di Bernado, D. (2006), Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, *Bioinformatics* **22(7)**, 815-822.

[Becker *et al.*, 2002] Becker, B., Hereth, J., Stumme, G. (2002) ToscanaJ: An open source tool for qualitative data analysis, *Advances in Formal Concept Analysis for Knowledge Discovery in Databases*, 1-2.

[Boshoff *et al.*, 2004] Boshoff, H., Myers, T.G., Copp, B.R., McNeill, M.R., Wilson, M.A., Barry, C.E. (2004) The transcriptional response of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action, *Biol.Chem.* **279(38)**, 40174-84.

[Cabusora *et al.*, 2005] Cabusora, L., Sutton, E., Fulmer, A., Forst, C.V. (2005) Differential network expression during drug and stress response, *Bioinformatics* **21**, 2898-2905.

[Dullaghan *et al.*, 2002] Dullaghan, E.M., Brooks, P.C., Davis, E.O.(2002) The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene - identification of a novel DNA-damage-inducible gene, *Microbiology* **148**, 3609-3615.

[Ganter and Wille, 1996] Ganter, B., Wille, R. (1997) Formal concept analysis, Mathematical Foundations, *Springer-Verlag.*

[Gebert *et al.*, 2006] Gebert, J., Radde, N. (2006) A new approach for modeling procaryotic biochemical networks with differential equations, *AIP Conference Proceedings* **839**, 526-533.

[Hashimoto *et al.*, 2004] Hashimoto, R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., Dougherty, E.R. (2004) Growing genetic regulatory networks from seed genes, *Bioinformatics* **20**, 1241-1247.

[Ihmels *et al.*, 2002] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network, *Nature Genetics* **31**, 370-377.

[Kloster *et al.*, 2005] Kloster, M., Tang, C., Wingreen, N.S. (2005) Finding regulatory modules through large-scale gene-expression data analysis, *Bioinformatics* **21(7)**, 1172-1179.

[Mawuenyega *et al.*, 2005] Mawuenyega, K.G., Forst, C.V., Dobos, K.M., Belisle, J.T., Chen, J., Bradbury, E.M., Bradbury, A.R.M., Chen, X. (2005) *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling, *Molecular Biology of the Cell* **16**, 396-404.

[Radde *et al.*, 2006] Radde, N., Gebert, J., Forst, C.V. (2006) Systematic component selection for gene-network refinement, *Bioinformatics* **22(21)**, 2674-2680.

[Troyanskaya *et al.*, 2001] Troyanskaya, O., Cantor, O., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics* **17(6)**, 520-525.

[Wille, 1982] Wille, R. (1982) Restructuring lattice theory: an approach based on hierarchies of concepts, *Ordered sets, editor Rival, I., publisher Reidel, Dordrecht-Boston*, 445-470.